DOI:10.12158/j.2096-3203.2024.01.025

基于相关性分析和生成对抗网络的电网缺失数据填补方法

蔡榕[1]，杨雪[2]，田江[1]，赵奇[1]，王毅[3]
（1. 国网江苏省电力有限公司苏州供电分公司，江苏 苏州 215004；
2. 南京工程学院创新创业学院，江苏 南京 211167；
3. 国网电力科学研究院有限公司，江苏 南京 211106）

摘 要：城市电网新型电力系统中多元资源增多，数据采集难度加大，导致数据随机缺失率升高，难以满足精细化分析决策需求。为解决新型电力系统中配网量测数据在采集与传输过程中频发的缺失问题，文中提出一种基于波动互相关分析（fluctuation cross-correlation analysis，FCCA）算法和生成对抗网络（generative adversarial network，GAN）的电网缺失数据填补方法。首先，融合 FCCA 算法提出强相关性电网数据多维特征提取方法；其次，基于核主成分分析（kernel principal component analysis，KPCA）对多维特征数据集进行降维处理；最后，设计改进型 GAN 结构，融合电网数据多维特征对低维向量进行重构，实现缺失数据填补。算例采用真实电网数据进行算法验证，并在某城市电网试运行。结果表明，所提方法比传统数据填补方法具有更高填补精度。因此，在新型电力系统中量测数据连续缺失和缺失量较大的情况下，融合强相关性特征进行数据填补，对提升量测数据的完整性和可用性有明显优势。

关键词：新型电力系统；波动互相关分析（FCCA）；多维特征；生成对抗网络（GAN）；缺失数据；核主成分分析（KPCA）；智能填补

中图分类号：TM714 **文献标志码**：A **文章编号**：2096-3203（2024）01-0229-09

0 引言

随着以新能源为主体的新型电力系统建设加快推进，地区电网作为能源使用的重要主体，已成为新型电力系统的主要建设方向。新型电力系统呈现新能源高比例渗透、电力电子设备规模化应用、分布式可调资源广泛接入、设备异构化、电网多形态等新特征[1-2]。融合系统中多元资源产生的海量多类型量测数据[3-4]对电力系统状态估计、设备评估、事故分析等具有重要意义[5-7]，因此在新型电力系统中保证电网设备量测数据的完整性和可用性尤为重要。

但在新型电力系统中，设备类型多、分布广、差异大，导致电网量测数据采集不确定性增加，数据随机缺失率升高，难以满足城市电网对电力数据进行精细化分析决策的需求。增加采集装置、调整采集频率、提高信道质量、优化通信机制等方法虽能解决数据缺失问题，但其综合成本高、建设周期长、施工难度大。因此，以采集的量测数据为研究对象，开展缺失数据填补方法研究，提升电网数据完整性和可用性成为了电力系统领域的一个重要研究方向[8-11]。

收稿日期：2023-09-15；修回日期：2023-11-26
基金项目：国家电网有限公司总部科技项目（5108-2022182-80A-2-296-XG）

缺失数据填补源于统计学分析方法，k 近邻（k-nearest neighbor，KNN）算法、随机森林算法、反向传播（back propagation，BP）神经网络算法等都较适用于缺失数据填补领域[12-15]。在电力系统范畴，长短期记忆网络和生成对抗网络（generative adversarial network，GAN）被用于重构电力系统量测缺失数据[16-18]。针对电网中的光伏数据，也可利用 GAN 处理缺失数据，以取得较好的填补效果[19-21]。但由于新型电力系统中强随机性、波动性的新能源大规模并网，电动汽车、分布式电源等交互式设备大量接入，以及噪声干扰、网络延迟等复杂因素的影响，电网数据特征类型更加多样化，缺失数据分布范围更大。上述研究主要基于缺失数据本身的特征，很少利用多种相关性特征解决数据缺失问题。但新型电力系统中的多元化电力设备存在时间、地域、采样周期等数据差异，电压、谐波、有功功率、无功功率、电量等电网多维特征之间关联性更加复杂紧密，仅根据缺失数据本身的单特征进行填补不能保证数据的完整性和真实性，从而影响电力系统精细化分析、准确判断、新能源消纳评估和负荷调节能力，难以满足城市电网高质量发展需求。

在此背景下，为面对新型电力系统环境下量测数据缺失给基础数据预处理和电力数据精细化分析决策带来的新挑战，文中针对配网管理系统（distribution management system，DMS）中的量测数

据预处理问题，提出了一种基于相关性分析的Wasserstein距离生成对抗网络（Wasserstein generative adversarial network based on correlation analysis，WGAN-CA）的电网缺失数据填补方法。首先，融合波动互相关分析（fluctuation cross-correlation analysis，FCCA）算法，提出强相关性电网数据多维特征提取方法，获得与缺失数据关联性较强的多维特征；其次，基于核主成分分析（kernel principal component analysis，KPCA）算法设计特征数据集降维映射方法，为数据填补网络提供输入数据；最后，提出WGAN-CA的网络结构，拟合数据多维特征对低维向量进行重构，填补电网缺失数据。算例采用真实电网数据进行算法验证，并在某城市电网试运行，结果表明所提方法具有更高的数据填补精度，能有效提升量测数据的完整性和可用性，更适合新型电力系统复杂数据环境，为电力系统高效运行及故障排除提供高质量数据基础。

1 融合FCCA的电网数据多维特征提取方法

FCCA可以将波动分析算法扩展到2个时间序列中[22]，从整体寻找2个序列之间的互相关联性，从而获得新型电力系统中各种量测数据间的相关性，为选取具有强相关的特征数据集提供依据。

若电网中某2种类型的设备量测数据为 x 和 y，每种数据在一个采样周期内进行了 n 次采样，表示为 $x=\{x_1,x_2,\cdots,x_n\}$，$y=\{y_1,y_2,\cdots,y_n\}$。首先计算序列的离差并求和，x 和 y 序列的离差求和分别为 $\Delta x(l)$ 和 $\Delta y(l)$。

$$\Delta x(l)=\sum_{i=1}^{l}(x_i-\bar{x}) \tag{1}$$

$$\Delta y(l)=\sum_{i=1}^{l}(y_i-\bar{y}) \tag{2}$$

式中：$l=1,2,\cdots,n$；x_i、y_i 分别为2个序列中第 i 个时间点的量测值；$\bar{x}$、$\bar{y}$ 分别为2个序列所有量测数据的平均值。

然后，计算代表时间序列自相关性的前向差分：

$$\Delta x(l,l_0)=x(l_0+l)-x(l_0) \tag{3}$$

$$\Delta y(l,l_0)=y(l_0+l)-y(l_0) \tag{4}$$

式中：$l_0=1,2,\cdots,n-1$。

最后，计算序列的协方差：

$$\mathrm{Cov}_{x,y}(l)=\sqrt{(\Delta x(l,l_0)-\overline{\Delta x(l,l_0)})(\Delta y(l,l_0)-\overline{\Delta y(l,l_0)})} \tag{5}$$

式中：$\overline{\Delta x(l,l_0)}$、$\overline{\Delta y(l,l_0)}$ 分别为2个序列前向差分的平均值。

从式(5)中可以看出，若2个序列存在关联性，$\mathrm{Cov}_{x,y}(l)$ 服从幂律分布，即 $\mathrm{Cov}_{x,y}(l)\sim l^{h_{x,y}}$，其中 $h_{x,y}$ 为序列 x 与 y 的关联系数。当 $h_{x,y}=0$ 时，序列 x 与 y 无关；当 $h_{x,y}<0$ 时，序列 x 与 y 呈现负相关；当 $h_{x,y}>0$ 时，序列 x 与 y 呈现正相关。$h_{x,y}$ 数值越大，2个序列之间的相关性越强。

融合FCCA可以计算电力系统多种类型量测数据序列之间的关联系数，从而选取具有强相关性的数据类型作为多维特征，开展缺失数据填补分析。设有 N 种类型的电网量测数据（如电压、电流、有功功率等），分别表示为 D_1、D_2、…、D_N。采用FCCA算法分别计算 D_1 与 D_2—D_N 之间的关联系数，即根据式(5)计算 h_{D_1,D_2}—h_{D_1,D_N}，并将 h_{D_1,D_2}—h_{D_1,D_N} 按降序排列。最后根据精度要求确定关联系数阈值，大于阈值的数据类型被提取为与 D_1 有强相关性的多维特征。

2 核主成分分析（KPCA）

KPCA不仅能降低数据集的维度，还可以挖掘出数据集中蕴含的非线性特征，因此常被用于非线性数据集的降维处理[23]。KPCA降维过程中，对于输入空间中的矩阵 $\boldsymbol{X}$，先用一个非线性映射将 $\boldsymbol{X}$ 中的所有样本映射到一个高维甚至是无穷维的空间，使其线性可分；然后在这个高维空间进行主成分分析（principal component analysis，PCA）降维，提取主成分。此过程既可以保留电网设备多种量测原始数据间的非线性关系，又可以降低数据维度。

若电网量测数据样本存储为矩阵 $\boldsymbol{X}$，$\boldsymbol{X}$ 的每一列表示一个样本，即 $\boldsymbol{X}=[\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_M]$。每个样本点为 $\boldsymbol{x}_i$，其为一个 R_d 维向量。根据KPCA的思想，用非线性映射 $\boldsymbol{\Phi}$ 将 $\boldsymbol{X}$ 中的向量 $\boldsymbol{x}_i$ 映射到高维空间 F 下。将 $\boldsymbol{X}$ 中所有样本都映射到 F，得到新矩阵 $\boldsymbol{\Phi}(\boldsymbol{X})=[\boldsymbol{\Phi}(\boldsymbol{x}_1)\ \boldsymbol{\Phi}(\boldsymbol{x}_2)\ \cdots\ \boldsymbol{\Phi}(\boldsymbol{x}_M)]$。特征空间数据满足中心化的条件为：

$$\sum_{\mu=1}^{M}\boldsymbol{\Phi}(\boldsymbol{x}_\mu)=\boldsymbol{0} \tag{6}$$

式中：$\boldsymbol{\Phi}(\boldsymbol{x}_\mu)$ 为 $\boldsymbol{x}_\mu$ 映射到高维空间的样本，$\boldsymbol{x}_\mu$ 为满足中心化条件的样本点；M 为参考数据集的样本个数。

则特征空间的协方差矩阵 $\boldsymbol{C}$ 为：

$$\boldsymbol{C}=\frac{1}{M}\sum_{\mu=1}^{M}\boldsymbol{\Phi}(\boldsymbol{x}_\mu)\boldsymbol{\Phi}(\boldsymbol{x}_\mu)^{\mathrm{T}} \tag{7}$$

设 $\boldsymbol{C}$ 的特征值为 λ，特征向量为 $\boldsymbol{v}$，则可得：

$$\boldsymbol{C}\boldsymbol{v}=\lambda\boldsymbol{v} \tag{8}$$

即有:

$$\boldsymbol{\Phi}(\boldsymbol{x}_\mu)\times\boldsymbol{v}=\lambda(\boldsymbol{\Phi}(\boldsymbol{x}_\mu)\times\boldsymbol{v}) \tag{9}$$

通过计算得到合适的系数 α_i ,可将所有特征向量表示为 $\boldsymbol{\Phi}(\boldsymbol{x}_1)$、$\boldsymbol{\Phi}(\boldsymbol{x}_2)$、…、$\boldsymbol{\Phi}(\boldsymbol{x}_M)$的线性张成:

$$\boldsymbol{v}=\sum_{i=1}^{M}\alpha_i\boldsymbol{\Phi}(\boldsymbol{x}_i) \tag{10}$$

结合式(7)—式(10)则有:

$$\frac{1}{M}\sum_{\mu=1}^{M}\sum_{i=1}^{M}\left[\boldsymbol{\Phi}(\boldsymbol{x}_\mu)\boldsymbol{\Phi}(\boldsymbol{x}_\mu)^{\mathrm{T}}(\alpha_i\boldsymbol{\Phi}(\boldsymbol{x}_i)\boldsymbol{\Phi}(\boldsymbol{x}_k))\right]=\lambda\left(\sum_{i=1}^{M}\alpha_i\boldsymbol{\Phi}(\boldsymbol{x}_i)\boldsymbol{\Phi}(\boldsymbol{x}_k)\right) \tag{11}$$

式中:$k=1,2,\cdots,M$。定义 $M\times M$ 维核矩阵 $\boldsymbol{K}$ 为:

$$\boldsymbol{K}=(\boldsymbol{\Phi}(\boldsymbol{x}_i)\boldsymbol{\Phi}(\boldsymbol{x}_\mu)) \tag{12}$$

求解得到核矩阵的特征值及特征向量,将其作为电网量测数据样本 D_1、D_2、…、D_N 构成的输入空间在特征空间的投影。经过 KPCA 算法处理后,电网量测数据集可以被映射为一个低维向量,便于进一步分析与辨识。

3 基于多维特征分析 WGAN-CA 的缺失电网数据填补

3.1 WGAN-CA 网络结构

GAN 自提出以来,在数据生成方面优势极大,已经被普遍用于图像生成、文本生成等领域[24]。考虑到缺失数据生成与图像生成在本质上是一致的,都是通过生成器与判别器之间的相互学习、博弈,生成约束条件下的缺失数据,故基于 GAN 结构开展电网缺失数据填补方法研究。但 GAN 模型存在训练不稳定及模式崩溃问题[25],而 WGAN 模型采用 Wasserstein距离[26-27]优化目标能够使训练过程更加稳定,达到更好的训练效果。因此,对于新型电力系统中电网缺失数据问题,融合电网数据多维特征,对生成器和判别器进行结构优化,并提出了 WGAN-CA 网络结构,如图 1 所示。

WGAN-CA 网络主要包括生成器和判别器两部分。在生成器中,输入随机噪声数据,通过全连接层单元对输入数据进行预测,生成一系列电网设备量测数据。在判别器中,首先通过相关性分析选出与待填补数据具有强相关性的样本类型,再将这些数据和生成器生成的数据一并输入判别器。经判别器中 4 个全连接层的特征分析后,得到关于生成数据真实性的判别结果。判定为真实的数据会被输出,同时也会成为网络参数优化的依据。

在 WGAN-CA 中,考虑到新型电力系统中各种配网设备量测数据多为一维时序数据,为保证生成

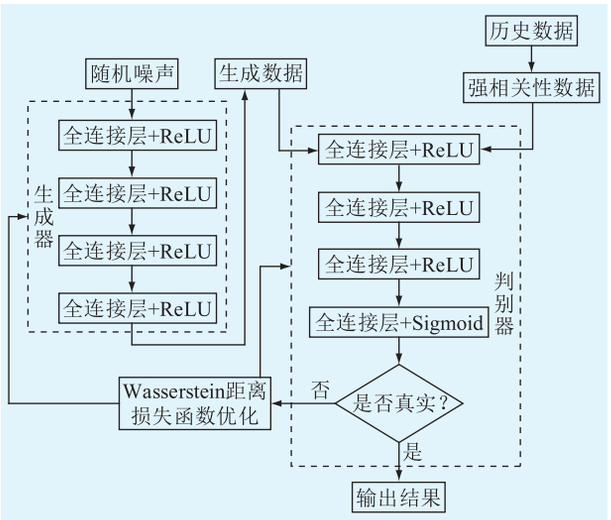


图 1 WGAN-CA 网络结构

Fig.1 WGAN-CA network structure

数据序列的精度和网络的收敛速度,根据该类型数据的数据量、数据结构等特征,设定网络层数、神经元个数及激活函数。生成器采用 4 层一维全连接层结构,如表 1 所示。每个全连接层的激活函数均采用 ReLU,以提高网络训练的效率。判别器的网络结构如表 2 所示,判别器与生成器的网络参数及结构基本一致,唯一的不同点是最后一层全连接层的激活函数为 Sigmoid。这是由于判别器最后的输出必须是一个概率值,用来判断生成数据的真实性,因此使用 Sigmoid 函数来提高填补数据的精度。

表 1 生成器网络结构

Table 1 Generator network structure

层数	网络名称	神经元个数	激活函数
1	全连接层	32	ReLU
2	全连接层	64	ReLU
3	全连接层	64	ReLU
4	全连接层	32	ReLU

表 2 判别器网络结构

Table 2 Discriminator network structure

层数	网络名称	神经元个数	激活函数
1	全连接层	32	ReLU
2	全连接层	64	ReLU
3	全连接层	16	ReLU
4	全连接层	1	Sigmoid

WGAN-CA 中,生成器和判别器的目标函数 $V(D,G)$ 均采用最大最小目标函数:

$$\min_G\max_D V(D,G)=E_{r\sim p_1(r)}(\lg D(r))+E_{z\sim p_z(z)}(\lg(1-D(G(z)))) \tag{13}$$

式中:z 为随机噪声;r 为真实数据;$D(r)$ 为判别器对

于真实数据的输出；$G(z)$ 为生成器对于噪声样本的输出；$p_t(r)$ 为真实数据分布；$p_z(z)$ 为高斯随机噪声分布；$E_{r\sim p_t(r)}$ 为 r 关于 $p_t(r)$ 的数学期望；$E_{z\sim p_z(z)}$ 为 z 关于 $p_z(z)$ 的数学期望。

$\min\limits_{G}\max\limits_{D}V(D,G)$ 表示判别器的训练目标是函数 V 取值最大化，而生成器的训练目标是函数 V 取值最小化。在训练过程中，通过调整生成器的参数使 $\lg(1-D(G(z)))$ 最小化；通过调整判别器的参数使 $\lg D(r)$ 最小化。生成器和判别器之间的对抗过程持续不断，直至达到平衡，生成器不能再被训练和优化。

在优化过程中，为了提高训练进程指标的可靠性，同时解决训练不稳定的问题，WGAN-CA 使用 Wasserstein 距离损失函数 $W(P_g,P_r)$ 优化训练进程，不断调整生成器和判别器中的网络参数。

$$W(P_g,P_r)=\inf_{\gamma\sim\prod(P_g,P_r)}E_{(a,b)\sim\gamma}(\|a-b\|) \tag{14}$$

式中：$W(P_g,P_r)$ 为将 P_g 拟合到 P_r 需要将 a 移动到 b 的距离；P_g、P_r 分别为生成数据序列和真实数据序列；a、b 为 2 个二维空间中的均匀分布；$E_{(a,b)\sim\gamma}$ 为由 a 移动到 b 从而让 a、b 服从相同分布 γ 的数学期望；$\prod(P_g,P_r)$ 为以 P_g 和 P_r 为边缘分布的联合概率分布 γ 的集合。$W(P_g,P_r)$ 数值越小，真实数据序列分布与生成数据序列分布越接近，训练效果就越好。生成器和判别器之间不断对抗优化，直至达到平衡，最终生成符合实际分布的电网量测数据序列。

3.2 基于相关性分析的多维特征提取

为消除数据之间不同量纲对模型训练效率的影响，训练数据输入网络前要进行标准化处理。Z-score标准化能有效避免处理结果较为接近且容易趋于 0 的问题，使标准化的数据处于标准正态分布中，突出不同特征对缺失数据的影响，有利于提高特征选择效果[28]。采用式(15)对采集的电网设备原始数据进行 Z-score 标准化处理。

$$\hat{m}_x=\frac{m_x-a(m_x)}{s_{td}} \tag{15}$$

式中：m_x 为某种类型的设备量测数据值；$\hat{m}_x$ 为标准化后的数值；$a(m_x)$ 为 m_x 对应特征的平均值；s_{td} 为 m_x 对应特征所有数的标准差。

与传统电力系统相比，新型电力系统的电网数据特征类型更加多样化，电网数据特征也各不相同。若将电网数据直接送入网络训练，训练效率较低且容易欠拟合。因此，将筛选出的与所分析数据类型具有强相关性的多维特征作为输入，可以有效提高数据填补精度。根据融合 FCCA 的电网数据多维特征提取方法，可提取与缺失数据类型高度相关的多维特征，作为缺失数据填补网络的输入。

3.3 基于 WGAN-CA 的数据填补策略

为正确填补电网系统中设备量测缺失数据，文中设计了基于 WGAN-CA 的数据填补策略，填补流程如图 2 所示。

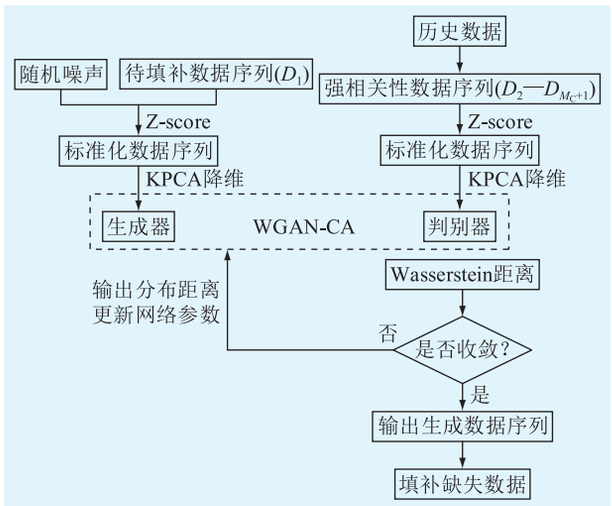


图 2 数据填补流程

Fig.2 Flow chart of data filling

训练判别器时，首先，生成一个符合高斯分布的随机噪声来填补缺失数据，与真实数据构成训练数据集。接着，基于 FCCA 从历史数据中获得待填补数据的强相关性数据序列。然后，为提升网络的训练效率，采用 KPCA 算法对训练集进行降维处理，将降维后的训练集输入生成器生成数据，并根据目标数据得出判别器的损失值。最后，采用 RMSProp 优化器减小损失值，不断迭代更新网络权重。训练生成器时，停止判别器更新网络权重，采用 RMSProp 更新网络权重并计算生成器的损失值，更新网络权重。为了提高训练效率，要按照先更新判别器网络，再更新生成器网络的顺序。当 WGAN-CA 模型训练到生成器和判别器的损失函数基本稳定，损失值无明显变化时，可将其视为已经收敛的模型，用于生成数据序列。

设 N 种类型的电网量测数据(如电压、电流、有功功率等)中有一个类型(如电压)的数据序列 D_1 存在缺失数据，首先，基于 FCCA 筛选出与该类型有强相关性的 M_C 个数据类型(如电流、有功功率等)的序列 D_2—D_{M_C+1}；接着，对 D_1—D_{M_C+1} 进行标准化处理；然后，通过 KPCA 将其降维处理后共同输入 WGAN-CA，经判别器和生成器不断训练 Wasserstein 距离损失函数；最后，生成与 D_1 真实数据序列一致

的数据序列并输出,完成对 D_1 中缺失数据的填补。

4 算例分析

选取华东某地级市电网中分布式电源接入点一个季度采集的数据作为算例,采样频率为 10 min,共计 12 960 组数据样本。WGAN-CA 采用 Pytorch 深度学习框架并按照表 1、表 2 进行网络搭建。由于篇幅所限,以电网电压数据填补为例,选取 GAN、KNN、随机森林和样条插值算法与文中所提方法进行对比分析。

4.1 误差分析标准

为了量化缺失数据填补效果,文中采用平均执行时间(mean execution time,mET)为数据填补速度评价指标,采用均方根误差 E_{rmse}、平均绝对误差 E_{mae} 和偏差率 E 为数据填补模型的精度评价指标。各指标计算如式(16)—式(18)所示,其中 x_i^* 为填补值。运算时间越短,误差值和偏差率越小,模型性能越好。

$$E_{\mathrm{rmse}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-x_i^*)^2} \tag{16}$$

$$E_{\mathrm{mae}}=\frac{1}{n}\sum_{i=1}^{n}|x_i-x_i^*| \tag{17}$$

$$E=\frac{|x_i-x_i^*|}{x_i}\times 100\% \tag{18}$$

4.2 多维特征提取与降维融合效果测试

算例数据包含电压、电流、有功功率、无功功率、电阻、电抗、功率因数和电量等 8 种数据类型,以填补电压缺失数据为例,采用 FCCA 计算电压与各个特征的互相关系数,结果如表 3 所示。

表 3 FCCA 算法计算结果

Table 3 Calculation results of FCCA algorithm

特征	互相关系数	特征	互相关系数
电流	0.72	电抗	0.80
有功功率	0.79	功率因数	0.78
无功功率	0.85	电量	0.79
电阻	0.78		

由表 3 可知,FCCA 算法充分挖掘了电压与其他特征之间隐藏的非线性关系,波动互相关系数越高的特征与电压关联越紧密。表 4 为不同 WGAN-CA 输入维度下电压缺失数据填补误差对比结果。

由表 4 可知,输入维度为 5 时,WGAN-CA 的填补误差最小,填补效果最好。输入从 3 维增加到 4 维时,输入网络的有效信息量增加,填补精度略微升高;输入从 5 维增加到 7 维时,输入网络的冗余信息量增加,填补误差有所增加。因此,将关联系数阈值设置为 0.79 时,有功功率、无功功率、电抗、电量 4 个特征与电压的相关性最强,将其作为多维度特征变量输入 WGAN-CA 网络进行缺失数据填补。

表 4 不同输入维度下的填补误差对比

Table 4 Comparison of filling errors under different input dimensions

输入维度	E_{rmse}/%	E_{mae}/%
3	1.95	1.10
4	1.89	1.09
5	1.78	0.93
6	1.85	1.05
7	1.82	1.11

同时,为测试融合 FCCA 在数据填补中的优势,将所提 WGAN-CA 与传统 WGAN 的填补效果进行对比。以相同数据集为例,将电压、电流、有功功率、无功功率等上述 8 种特征作为输入类型。对于 WGAN,因未进行 FCCA 处理,8 种特征同时输入,此时网络结构为 8 个输入维度。而对于 WGAN-CA,经 FCCA 后选择 4 个强相关性特征和电压特征作为输入,此时网络结构为 5 个输入维度。基于 2 种网络对电压缺失数据进行填补的误差与运算时间如表 5 所示,对比可知融合相关性分析可获得更高的填补精度和更快的运算速度。

表 5 融合相关性分析对填补效果的影响对比

Table 5 Comparison of correlation analysis fusion impact on filling effect

网络名称	E_{rmse}/%	E_{mae}/%	mET/s
WGAN	1.89	1.10	5.72
WGAN-CA	1.78	0.93	1.29

另一方面,从表 3 中易发现,电阻、功率因数与电压之间同样具有较高的波动互相关系数。在实验中与电压相关性较高的特征都应该加以考虑,因此将关联系数阈值调整为 0.78。此时,有功功率、无功功率、电阻、电抗、功率因数和电量 6 个特征被确定为电压的强相关性特征。将其与电压作为 WGAN-CA 的输入时,网络结构要调整为 7 个输入维度。这样虽然关联因素更多,但网络结构的复杂程度也随之增加。为此,对筛选出的 7 维数据进行标准化处理后,使用 KPCA 算法对其进行降维处理。

由于电网设备量测数据可被视为没有先验知识的非线性数据,所以采用高斯径向基函数作为 KPCA 核函数,主元贡献率阈值设置为 80%。经测试,当主元个数为 5 时,主元贡献率为 80.7%,已达到要求。因此经 KPCA 处理后,将 7 维数据映射到

一个 5 维度的空间中，作为 WGAN-CA 网络的输入。有无 KPCA 降维处理对电压缺失数据的填补效果对比如表 6 所示。

表 6　KPCA 降维对填补效果的影响

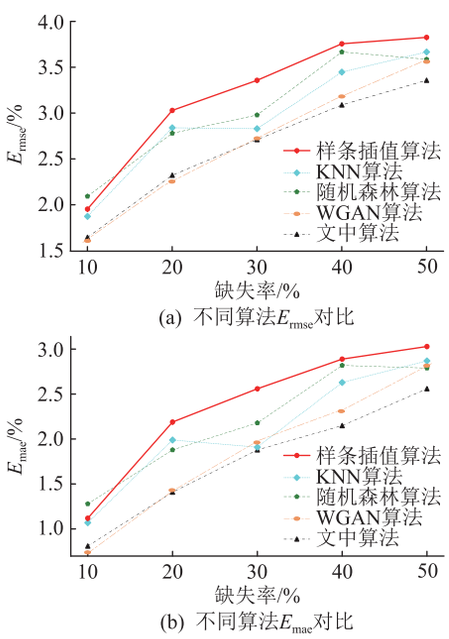
Table 6　KPCA dimensionality reduction impact on filling effect

输入维度	有无 KPCA 降维	E_{rmse}/%	E_{mae}/%	mET/s
5	无	1.78	0.93	1.29
7	无	1.82	1.11	2.53
5	有	1.29	0.79	1.28

经对比可见，采用 KPCA 降维处理既符合了该网络的最优输入维度要求，又保留了输入数据中重要的相关性信息。所以，文中提出的 WGAN-CA 算法充分融合了与电压具有强相关性的特征进行缺失数据填补，不仅提高了数据填补精度，还能有效精简网络结构，缩短执行时间。

4.3　填补结果分析

从电网数据中筛选完整的子序列，按照随机删除的方法分别构建比例为 10%、20%、30%、40%、50%的缺失数据，在不同缺失比例下分别应用文中算法（WGAN-CA）、WGAN、KNN、随机森林和样条插值 5 种算法进行仿真分析，填补结果精度见图 3。



(a) 不同算法E_{rmse}对比

(b) 不同算法E_{mae}对比

图 3　5 种算法填补精度对比

Fig.3　Filling accuracy comparison of five algorithms

由图 3 可知，5 种算法填补数据的 E_{rmse} 和 E_{mae} 变化趋势基本相同，文中算法在各种缺失比例条件下 E_{rmse} 和 E_{mae} 均最小，填补精度最高。虽然 WGAN 算法在缺失率非常低的情况下填补精度比文中算法略高，但随着缺失率的增大，该方法的数据填补精度大幅降低。另一方面，文中算法融合了相关性分析，在运算时间上也明显优于 WGAN。

为测试文中算法的数据填补精度，将 12 960 组数据中新能源设备高比例渗透的样本数据筛选出来，从中随机选择包含 1 000 个电压数据的连续序列，删除部分数据后进行缺失数据填补精度测试。分别测试缺失率为 10%、20%、30%、40%和 50%的数据填补情况。当填补值与真实值间的偏差率小于 1%时，认为填补正确，否则填补错误。5 种算法的缺失数据填补准确率综合对比见图 4。

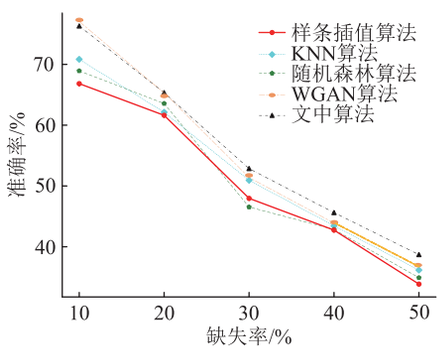


图 4　不同算法的填补准确率对比

Fig.4　Comparison of filling accuracy of different algorithms

在数据缺失率为 20%～30%时，5 种算法填补准确率差距较小；但文中算法的填补准确率仍比其余 4 种算法高约 4.8%～8.2%。当数据缺失率为 40%时，样条插值算法、KNN 算法和随机森林算法的填补准确率基本相同，文中算法虽然填补准确率也有所下降，但仍比另外 4 种算法高约 14.3%。当缺失率为 50%时，文中算法的填补准确率还比其他 4 种算法高约 11.5%。

另一方面，由于新型电力系统中电网数据特征类型更加多样化，缺失数据分布范围更大，缺失率明显高于传统电网数据。因此在实验测试中将 30 个数据的样本序列随机删除 9 个，采用 5 种算法对缺失率达 30%的数据集进行缺失填补测试。测试样本数据集如图 5 所示，填补结果对比如表 7 所示。

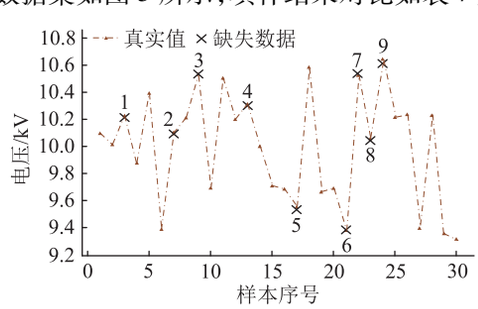


图 5　缺失数据样本分布

Fig.5　Missing data sample distribution

表 7 5 种算法填补误差率对比

Table 7 Comparison of the filling errors of five algorithms

填补算法	填补偏差率/%									平均偏差率/%
	样本 1	样本 2	样本 3	样本 4	样本 5	样本 6	样本 7	样本 8	样本 9	
WGAN-CA	0.98	0.83	0.97	0.65	0.56	1.83	2.13	2.31	1.07	1.26
WGAN	0.99	0.85	1.02	0.52	0.49	1.97	2.33	2.65	1.13	1.33
KNN	1.03	0.92	1.99	1.46	0.72	1.67	3.45	4.36	2.07	1.96
随机森林	2.33	7.89	8.21	6.01	5.33	7.09	7.98	13.23	9.37	7.49
样条差值	3.32	7.67	7.66	5.31	4.92	6.56	8.21	12.55	8.32	7.17

从表 7 中可看出,文中算法的缺失数据填补平均偏差率最小。虽然在样本 4 和样本 5 中,WGAN 能达到略高的填补精度,但文中算法在平均精度和运算时间方面都明显优于 WGAN。另外,由于样本 6—样本 9 是连续缺失样本,因此针对其填补效果,5 种算法都存在较大误差,但文中算法的填补偏差率明显低于其余 4 种算法。因此,利用 WGAN-CA 算法对数据进行填补,填补结果与真实值更接近,能较好地还原缺失数据,更加符合数据填补要求。

同时,基于文中提出的关键技术,在江苏省某地级市电网调控系统平台进行了部署试运行,结合现场对数据采集与监视控制(supervisory control and data acquisition,SCADA)基础数据进行了多次缺失数据填补实效测试,经填补处理后的各种配网数据连续性和可用性均得到保证,数据填补效果良好。

5 结语

文中阐述了数据缺失的处理方法及相关理论,主要包括数据缺失的原因及处理的必要性,并介绍了相关的数据填补模型。针对新型电力系统多样化设备数据特征,为提高电网数据缺失值的填补准确度,文中提出了一种基于多维特征分析和 WGAN 的电网缺失数据智能填补方法。该方法通过分析缺失数据与设备其他类型数据相关性选取多维特征,提出 WGAN-CA 网络拟合电网设备数据多维特征对低维向量进行重构判别,填补电网缺失数据,提升新型电网设备量测数据的完整性和可靠性。算例采用真实电网数据对所提方法进行仿真验证,并在某城市电网试运行,结果表明所提方法与传统数据填补方法相比,在数据填补精度方面具有优势,特别是在数据连续缺失和缺失量较大时效果更佳,更适用于新型电力系统复杂的数据环境。

下一步研究将重点考虑优化 WGAN-CA 网络结构,针对不同应用场景下的数据特征,引入注意力机制自适应调整网络权重参数,进一步提高缺失数据填补方法的精度,扩大其适用范围。

参考文献:

[1] 任大伟,肖晋宇,侯金鸣,等. 双碳目标下我国新型电力系统的构建与演变研究[J]. 电网技术,2022,46(10):3831-3839.
REN Dawei,XIAO Jinyu,HOU Jinming,et al. Construction and evolution of China′s new power system under dual carbon goal [J]. Power System Technology,2022,46(10):3831-3839.

[2] 郝文斌,孟志高,张勇,等. 新型电力系统下多分布式电源接入配电网承载力评估方法研究[J]. 电力系统保护与控制,2023,51(14):23-33.
HAO Wenbin,MENG Zhigao,ZHANG Yong,et al. Carrying capacity evaluation of multiple distributed power supply access to the distribution network with the background of a new power system[J]. Power System Protection and Control, 2023, 51(14):23-33.

[3] 王琦,李峰,汤奕,等. 基于物理—数据融合模型的电网暂态频率特征在线预测方法[J]. 电力系统自动化,2018,42(19):1-9.
WANG Qi,LI Feng,TANG Yi,et al. On-line prediction method of transient frequency characteristics for power grid based on physical-statistical model[J]. Automation of Electric Power Systems,2018,42(19):1-9.

[4] 程书灿,赵彦普,张军飞,等. 电力设备多物理场仿真技术及软件发展现状[J]. 电力系统自动化,2022,46(10):121-137.
CHENG Shucan,ZHAO Yanpu,ZHANG Junfei,et al. State of the art of multiphysics simulation technology and software development for power equipment[J]. Automation of Electric Power Systems,2022,46(10):121-137.

[5] 马兴明,董成,毛新宇,等. 基于状态估计的海量多元异构智能电网数据压缩存储方法[J]. 电机与控制应用,2023,50(2):67-72,81.
MA Xingming,DONG Cheng,MAO Xinyu,et al. Data compression and storage method of massive multivariate heterogeneous smart grid based on state estimation[J]. Electric Machines & Control Application,2023,50(2):67-72,81.

[6] 马玉玲,李朝祥,曹中枢,等. 基于数据融合技术的电力系统鲁棒动态状态估计方法[J]. 智慧电力,2023,51(10):78-84.
MA Yuling,LI Chaoxiang,CAO Zhongshu,et al. Robust dynamic state estimation method for power systems based on data fusion[J]. Smart Power,2023,51(10):78-84.

[7] 张雷,王光华,李金铖,等. 大数据背景下考虑删失特点的继保设备运行状态评估[J]. 电力工程技术,2021,40(6):185-192.

ZHANG Lei, WANG Guanghua, LI Jinshuo, et al. Operating status assessment of protective equipment considering censored characteristics under background of big data[J]. Electric Power Engineering Technology, 2021, 40(6): 185-192.

[8] HEMANTH G R, CHARLES RAJA S. Proposing suitable data imputation methods by adopting a stage wise approach for various classes of smart meters missing data - practical approach [J]. Expert Systems with Applications, 2022, 187: 115911.

[9] 李绍坚,韦明超,甘静,等. 基于多维度相关性分析的电压缺失数据辨识方法研究[J]. 电气自动化,2021,43(1):63-66.
LI Shaojian, WEI Mingchao, GAN Jing, et al. Research on the identification method of voltage missing data based on multidimensional correlation analysis[J]. Electrical Automation, 2021, 43(1): 63-66.

[10] 梅玉杰,李勇,周王峰,等. 基于机器学习的配电网异常缺失数据动态清洗方法[J]. 电力系统保护与控制,2023,51(7):158-169.
MEI Yujie, LI Yong, ZHOU Wangfeng, et al. Dynamic data cleaning method of abnormal and missing data in a distribution network based on machine learning[J]. Power System Protection and Control, 2023, 51(7): 158-169.

[11] ZHU J E, XU W X. Real-time data filling and automatic retrieval algorithm of road traffic based on deep-learning method [J]. Symmetry, 2020, 13(1): 1.

[12] 蒋辉,马超群,许旭庆,等. 仿 EM 的多变量缺失数据填补算法及其在信用评估中的应用[J]. 中国管理科学,2019,27(3):11-19.
JIANG Hui, MA Chaoqun, XU Xuqing, et al. An EM-similar imputation algorithm for multivariable data missing and its application in credit scoring[J]. Chinese Journal of Management Science, 2019, 27(3): 11-19.

[13] 赵春霞,赵营颖. 基于多元回归 KNN 的网络数据库不完整信息填充[J]. 计算机仿真,2021,38(8):339-343.
ZHAO Chunxia, ZHAO Yingying. Incomplete information filling of network database based on multiple regression KNN[J]. Computer Simulation, 2021, 38(8): 339-343.

[14] 杨晔民,张慧军,张小龙. 随机森林的可解释性可视分析方法研究[J]. 计算机工程与应用,2021,57(6):168-175.
YANG Yemin, ZHANG Huijun, ZHANG Xiaolong. Research on interpretable visual analysis method of random forest[J]. Computer Engineering and Applications, 2021, 57(6): 168-175.

[15] 王婧骅,张娟,赵婉茹,等. 基于时间序列的分布式光伏电站发电数据采集方法[J]. 电网与清洁能源,2022,38(6):137-142.
WANG Jinghua, ZHANG Juan, ZHAO Wanru, et al. A method for collecting power generation data of distributed photovoltaic power station based on time series[J]. Power System and Clean Energy, 2022, 38(6): 137-142.

[16] 王守相,陈海文,潘志新,等. 采用改进生成式对抗网络的电力系统量测缺失数据重建方法[J]. 中国电机工程学报,2019,39(1):56-64,320.
WANG Shouxiang, CHEN Haiwen, PAN Zhixin, et al. A reconstruction method for missing data in power system measurement using an improved generative adversarial network[J]. Proceedings of the CSEE, 2019, 39(1): 56-64, 320.

[17] 王子馨,胡俊杰,刘宝柱. 基于长短期记忆网络的电力系统量测缺失数据恢复方法[J]. 电力建设,2021,42(5):1-8.
WANG Zixin, HU Junjie, LIU Baozhu. Recovery method for missing measurement data of power systems based on long short-term memory networks[J]. Electric Power Construction, 2021, 42(5): 1-8.

[18] 修春波,苏欢,苏雪苗. 多通道长短期记忆卷积网络的风速预测[J]. 电力工程技术,2022,41(1):64-69.
XIU Chunbo, SU Huan, SU Xuemiao. Wind speed prediction based on multi-channel long short-term memory convolution neural network[J]. Electric Power Engineering Technology, 2022, 41(1): 64-69.

[19] 殷豪,丁伟锋,陈顺,等. 基于生成对抗网络和纵横交叉粒子群算法的光伏数据缺失重构方法[J]. 电网技术,2022,46(4):1372-1381.
YIN Hao, DING Weifeng, CHEN Shun, et al. Reconstruction method for missing data in photovoltaic based on generative adversarial network and crisscross particle swarm optimization algorithm[J]. Power System Technology, 2022, 46(4): 1372-1381.

[20] 卢俊菠,刘俊峰,罗燕,等. 基于改进 WGAN 考虑特征分布相似性的小样本负荷预测方法[J/OL]. 控制理论与应用: 1-11[2023-09-26]. http://kns.cnki.net/kcms/detail/44.1240.tp.20230324.1740.017.html.
LU Junbo, LIU Junfeng, LUO Yan, et al. Small sample load forecasting method considering characteristic distribution similarity based on improved WGAN[J/OL]. Control Theory & Applications: 1-11[2023-09-26]. http://kns.cnki.net/kcms/detail/44.1240.tp.20230324.1740.017.html.

[21] 李辉,任洲洋,胡博,等. 基于时序生成对抗网络的月度风光发电功率场景分析方法[J]. 中国电机工程学报,2022,42(2):537-548.
LI Hui, REN Zhouyang, HU Bo, et al. A sequential generative adversarial network based monthly scenario analysis method for wind and photovoltaic power[J]. Proceedings of the CSEE, 2022, 42(2): 537-548.

[22] 卢思安,侯国庆. 基于大数据分析技术的云计算资源预测研究[J]. 计算机仿真,2022,39(10):502-505,537.
LU Sian, HOU Guoqing. Research on cloud computing resource prediction based on big data analysis technology[J]. Computer Simulation, 2022, 39(10): 502-505, 537.

[23] 车昱娇,陈云霞,崔宇轩. KPCA 和改进 LSTM 在滚动轴承剩余寿命预测中的应用研究[J]. 电子测量与仪器学报,2021,35(2):109-114.
CHE Yujiao, CHEN Yunxia, CUI Yuxuan. Rolling element bearing remaining useful life estimation based on KPCA and improved long-short-term memory network[J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(2): 109-

114.

[24] 王德文,杨凯华. 基于生成式对抗网络的窃电检测数据生成方法[J]. 电网技术,2020,44(2):775-782.
WANG Dewen,YANG Kaihua. A data generation method for electricity theft detection using generative adversarial network [J]. Power System Technology,2020,44(2):775-782.

[25] BOUKRAICHI H, AKKARI N, CASENAVE F, et al. Uncertainty quantification in a mechanical submodel driven by a Wasserstein-GAN[J]. IFAC-PapersOnLine, 2022, 55(20): 469-474.

[26] 刘文斌,王兵,方刚,等. 基于中值的 JS 散度可变剪接差异分析研究[J]. 电子与信息学报,2020,42(6):1392-1400.
LIU Wenbin,WANG Bing,FANG Gang,et al. Study on the differential analysis of alternative splicing based on the median value Jensen-Shannon divergence[J]. Journal of Electronics & Information Technology,2020,42(6):1392-1400.

[27] 廖一帆,武志刚. 基于迁移学习与 Wasserstein 生成对抗网络的静态电压稳定临界样本生成方法[J]. 电网技术,2021,45(9):3722-3728.
LIAO Yifan, WU Zhigang. Critical sample generation method for static voltage stability based on transfer learning and Wasserstein generative adversarial network [J]. Power System Technology,2021,45(9):3722-3728.

[28] 任元秋,王兴,郑钦钦. 不同学科分类方案下不同学科标准化方法效果的比较研究[J]. 图书情报工作,2021,65(3):84-92.
REN Yuanqiu, WANG Xing, ZHENG Qinqin. Comparison of field normalization effects based on different discipline classification schemes[J]. Library and Information Service,2021,65(3):84-92.

作者简介:


蔡榕

蔡榕(1969),男,硕士,高级工程师,从事地区智能电网研究、规划、设计、建设、运行以及组织管理工作(E-mail:CR19692023@163.com);

杨雪(1982),女,硕士,副教授,研究方向为深度学习、人工智能;

田江(1981),男,硕士,高级工程师,从事电网调度自动化与智能化等相关工作。

A power system missing data filling method based on correlation analysis and generative adversarial network

CAI Rong[1], YANG Xue[2], TIAN Jiang[1], ZHAO Qi[1], WANG Yi[3]

(1. State Grid Suzhou Power Supply Company of Jiangsu Electric Power Co.,Ltd.,Suzhou 215004,China;
2. School of Innovation and Entrepreneurship,Nanjing Institute of Technology,Nanjing 211167,China;
3. State Grid Electric Power Research Institute Co.,Ltd.,Nanjing 211106,China)

Abstract: In the novel power system of urban grid, the multiple resources increase and the data collection becomes more difficult, which lead to a higher random missing data rate. It is difficult to meet the demand for refined analysis and decision making. For the frequent missing data problem in the distribution network, a new missing data filling method for power systems based on fluctuation cross-correlation analysis (FCCA) and generative adversarial network (GAN) is proposed in this paper. Firstly, a multi-dimensional feature extraction method for strongly correlated grid data is proposed by fusing FCCA. Secondly, based on kernel principal component analysis (KPCA), the multi-dimensional feature dataset is dimensionally reduced. Finally, an improved GAN structure is designed, which integrates multi-dimensional features of power grid equipment data to reconstruct low dimensional vectors. The missing data is accurately filled in, and the integrity and availability of the new power system measurement data is improved. The algorithm is validated using real grid data, and the proposed method is also tested in a city grid. The results show that the proposed method has higher filling accuracy than the traditional data filling methods. Therefore, it is conformed that in the case of continuous and significant data environment, integrating strong correlation features for data filling has significant advantages in improving the integrity and availability of measurement data.

Keywords: novel power systems; fluctuating cross-correlation analysis (FCCA); multi-dimensional features; generative adversarial networks (GAN); missing data; kernel principal component analysis (KPCA); intelligent filling

(编辑 陆海霞)